Investigating Stranded GMM for Improving Automatic Speech Recognition

Arseniy Gorin[1,2,3], *Denis Jouvet*[1,2,3], *Emmanuel Vincent*[1,2,3], *Dung Tran*[1,2,3]

Speech Group, LORIA
[1]Inria, 615 rue du Jardin Botanique, F-54600, Villers-lès-Nancy, France
[2]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[3]CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
{arseniy.gorin, denis.jouvet}@inria.fr

Abstract

This paper investigates recently proposed Stranded Gaussian Mixture acoustic Model (SGMM) for Automatic Speech Recognition (ASR). This model extends conventional hidden Markov model (HMM-GMM) by explicitly introducing dependencies between components of the observation Gaussian mixture densities.

The main objective of the paper is to experimentally study, how useful SGMM can be for dealing with data, which contains different sources of acoustic variability. First studied sources of variability are age and gender in quiet environment (TIdigits task including child speech). Second, the SGMM modeling is applied on data produced by different speakers and corrupted by non-stationary noise (CHiME 2013 challenge data). Finally, SGMM is applied on the same noisy data, but after performing speech enhancement (i.e., the remaining variability mostly comes from residual noise and different speakers).

Although SGMM was originally proposed for robust speech recognition of noisy data, in this work it was found, that the model is more efficient for handling speaker variability in quiet environment.

Index Terms: dynamic Bayesian network, hidden Markov model, trajectory modeling, robust speech recognition

1. Introduction

Hidden Markov models (HMMs) play an important role in statistical acoustic modeling for ASR. However, they are frequently criticized for inability to capture long temporal dependencies and variability of the speech signal because of strong conditional independence assumptions.

In the straightforward formulation it is assumed, that the state of an HMM is conditioned only on the previous state and that the observation is dependent only on the state which generates it. To better capture the contextual variability, context-dependent triphones are used. To capture speaker variability, the observation densities are built in the form of mixture of Gaussians. One problem, associated with conventional HMM is known as trajectory folding. This problem arises because of the fact, that the component of the GMM, trained from one source of variability, can dominate in likelihood computation for the observation from another source of variability [1].

Various techniques were introduced to relax the conditional independence assumptions and to achieve more accurate modeling of temporal dependencies. The classification of these techniques is done depending on where such dependencies are applied (in the feature, or in the model space) and on the type of such dependencies (linear or non-linear, with or without recursion) [2].

Most popular early approaches include Stochastic Segmental Models [3] and Stochastic Trajectory Models [4, 5]. Various approaches deal with additional dependencies in model space. These include Buried HMMs [6] with conditioned Gaussian observation densities and multi-path HMMs [7], where each separate path represents a separate source of variability. Recently, separately from HMM-GMM framework Deep Neural Networks-based systems have been successfully applied to better model both long temporal context and speech signal variability [8].

This work focuses on recently proposed Stranded Gaussian Mixture Model (SGMM) [9]. SGMM is an extension of HMM, which adds dependencies between components of HMM-GMM by expanding the observation density and replacing state-conditioned mixture weights by Mixture Transition Matrices (MTMs). MTM models transition probabilities between Gaussian components of the adjacent states.

In the original paper [9] SGMM was investigated on Aurora 2 connected digits task [10]. The authors used multi-condition training, i.e., including noised utterances in the training set. With standard MFCC front-end (12 cepstra + log energy, plus first and second derivatives) and Cepstral Mean Normalization (CMN) they demonstrated the improvement of average WER from 8.96% to 8.07% if GMM is replaced by SGMM with the same number of components. Such improvements encourage to further investigate the model for other sources of variability and on more difficult tasks.

Speaker age and gender are important sources of variability for ASR. Child speech is typically hard to recognize, because the acoustic variability comes from both vocal tract differences and specific articulation of children [11, 12]. If train and test sets are produced by both child and adult speakers, using all data to learn speaker-independent HMM-GMM leads to low performance. The variability can be partially reduced by using model and feature adaptation techniques, like MLLR, fMLLR [13] or VTLN [14]. However, adaptation requires prior knowledge about speaker classes (gender or age) and/or needs an additional pass in decoding for estimating such information.

Other sources of variability are the recording conditions (microphone, room reverberation) and noise. From this point of view, CHiME 2013 challenge [15] provides a hard task. Non-stationary noise in CHiME is added in random places of the utterances. In addition, the microphone movements are modeled. If only clean data is used for training, the accuracy of the model is very low. Better results are achieved when noised data is also used in training (multi-condition training). Speech en-

hancement [16] of both train and test set significantly improves the recognition accuracy, although the residual noise still makes the recognition task far from being considered as solved.

The experiments in this work are conducted using Sphinx3 toolkit [17], which was modified to handle SGMM. Speech enhancement is done using Flexible Audio Source Separation Toolbox (FASST) [18, 19].

The paper is organized as follows. Section 2 recaps the SGMM formulation. Section 3 describes the experiments with TIdigits task. Section 4 explains CHiME data and the corresponding experiments. The paper ends with a discussion and a conclusion.

2. Stranded GMM

The conventional SGMM consists of the state sequence $\mathcal{Q} = \{q_1, ..., q_T\}$, the observation sequence $\mathcal{O} = \{o_1, ..., o_T\}$, and the sequence of components of the observation density $\mathcal{M} = \{m_1, ..., m_T\}$, where every $m_t \in \{1, ..., M\}$ is the component of the observation density at the time t, and M denotes the number of such components in the mixture.

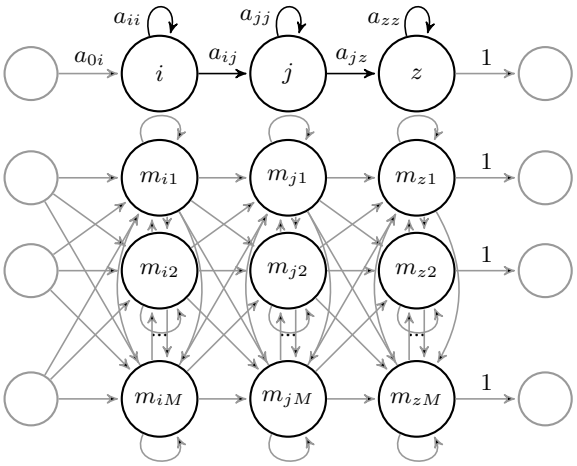

Figure 1: Stranded GMM for triphone HMM model

The difference of SGMM from HMM-GMM is that an additional dependency between the components of the mixture in the current frame m_t and at the previous frame m_{t-1} is introduced (Figure 1). The joint likelihood of the observation, state and component sequences is defined by:

$$\begin{aligned} p(\mathcal{O}, \mathcal{Q}, \mathcal{M}|\lambda) &= p(\mathcal{O}|\mathcal{M}, \mathcal{Q}, \lambda)p(\mathcal{M}|\mathcal{Q}, \lambda)p(\mathcal{Q}|\lambda) \\ &= \prod_{t=1}^{T} p(o_t|m_t, q_t)p(m_t|m_{t-1}, q_t, q_{t-1})p(q_t|q_{t-1}) \end{aligned} \quad (1)$$

The complete set of parameters is the following:

- $P(q_t = j|q_{t-1} = i) = a_{ij}$ - state transition probability;
- $P(o_t|m_t = l, q_t = j) = b_{jl}(o_t)$ - probability of the observation o_t with respect to the single density component $m_t = l$ in the state $q_t = j$;
- $P(m_t = l|m_{t-1} = k, q_t = j, q_{t-1} = i) = c_{kl}^{(ij)}$ - probability of moving from the component $m_{t-1} = k$ of the state $q_{t-1} = i$ to the component $m_t = l$ of the state $q_t = j$ (mixture transition probability)

The set of component transition probabilities forms the Mixture Transition Matrix (MTM) $C^{(ij)} = \{c_{kl}^{(ij)}\}$, where $\sum_{l=1}^{M} c_{kl}^{(ij)} = 1, \forall i, j, k$.

For training and decoding the Baum-Welch and Viterbi algorithms are expanded to include MTMs. The derivations can be found in [9].

3. SGMM for age and gender variability

To compare the SGMM behavior with different sources of variability, this section compares conventional HMM-GMM and Stranded HMM when dealing with clean data, which contains speakers of different age and gender.

3.1. TIdigits Baseline system and problem formulation

The experiments in this section are conducted on the TIdigits connected digits task [20]. The full training data set consists of 41224 digits (28329 for adult and 12895 for child speech). The test set consists of 41087 digits (28554 for adult and 12533 for child).

The digits are modeled as sequences of word-dependent phones. Each phone is modeled by a 3-state HMM without skips. Each state density is modeled by 32 Gaussian components. The front-end consists of standard 39 MFCC features (12 cepstra + log-energy, plus first and second order derivatives) with CMN.

Similar to other work with TIdigits [21], the signal is down sampled to 8 kHz. Word Error Rates (WERs) of the baseline systems are reported in Table 1. Two Speaker-Independent (SI) models are trained from the adult subset only and from the full training set. For the last two lines, the Age and Gender-Age dependent models are achieved with MLLR+MAP adaptation.

	Adult	Man	Wom	**Child**	Boy	Girl
Training on adult data	**0.64**	0.79	0.49	**9.92**	6.51	13.33
Training on adult+child data	**1.66**	1.86	1.46	**1.88**	1.69	2.08
+Age adaptation (classes known in decoding)	**1.42**	1.56	1.28	**1.56**	1.52	1.54
+Gender-Age adaptation (classes known in decoding)	**1.31**	1.57	1.04	**1.31**	1.14	1.49

Table 1: Baseline word error rates for SI, Age and Gender-Age adapted models with known speaker classes.

Training on adult data provides the best results for adult speakers, but shows a weak performance on child speech. When child data is included in the training set, the conventional HMM-GMM improves on the child, but degrades on the adult subset. Using class-adapted models further improves the baseline performance. In further experiments only full training set will be considered with no class information (i.e., age and gender) available.

3.2. Experimental results on TIdigits data

To train SGMM, MTM rows are initialized from the mixture weights of convention HMM-GMM, and the model parameters are re-estimated with MLE. In addition, to reduce the number of parameters, only 2 MTMs are used for each state (i.e., cross-phone MTMs are shared). Two SGMMs were built: with only MTMs and with all parameters re-estimated. The corresponding WERs are shown in Table 2.

Model	**Adult**	Man	Woman	**Child**	Boy	Girl
GMM	**1.66**	1.86	1.46	**1.88**	1.69	2.08
SGMM: MTM	**1.09**	1.22	0.96	**1.35**	1.30	1.41
SGMM: MTM+μ+σ	**1.11**	1.26	0.96	**1.27**	1.19	1.36

Table 2: Word error rates of stranded GMM vs conventional GMM on TIdigits task

Compared to the conventional HMM-GMM, SGMM improves from 1.66% to 1.11% on adult and from 1.88% to 1.27% on child speech. Both improvements are statistically significant with respect to 95% confidence interval. The SGMM performance is even better than the one achieved with the Gender-Age adapted baseline.

Notice, that the largest improvement is associated with only MTM re-estimation, whereas re-estimating means and variances only slightly improves on the child subset.

4. SGMM for robust speech recognition

This section analyzes the behavior of SGMM for noise-robust speech recognition on the 1^{st} track (small-vocabulary) of CHiME 2013 challenge.

4.1. CHiME task and baseline system

The task is to recognize digit and letter tokens in 6-word utterances. Overall there are 10 possible keywords for the digits and 25 keywords for the letters. The training set consists of 17000 utterances, which come from 34 different speakers (500 utterances per speaker). The development set is used for performance evaluation. The set contains 3600 utterances (600 for each SNR level). The utterances are corrupted by various types of non-stationary background noise with SNR from -6 to 9 dB.

In the baseline system each phone is modeled by an HMM with 3 states. However, the phones are not shared across different words (hence, word-dependent phones). Overall the model has 128 context-independent 3-state HMMs without skips. Each state is modeled by 32 Gaussian mixtures.

The performance is evaluated with 2 different types of acoustic features. In the first set of experiments standard 39 MFCC features (12 cepstra + log-energy, plus first and second order derivatives) with CMN are directly derived from noisy speech. In the second case, the same features are extracted after speech enhancement with uncertainty information, included in the feature computation (the approach is described in [16]). Speech enhancement allows to significantly reduce error rates compared to standard multi-condition training from noisy data.

In all experiments with SGMM, two MTMs for each model state are used, as in TIdigits experiments: one MTM for inter- and another for intra-state transition. The inter-state transitions between phone models are shared across different contexts. So, the total number of MTMs is equivalent to 2 times the number of states.

4.2. Mixture Transition Matrix analysis

It was observed, that inter- and intra-state MTMs are distributed differently. For example, the averaged values of MTMs for inter- and intra-state transitions computed with noisy data on CHiME dataset are presented in Figure 2. The averaged MTMs for TIdigits set has a similar appearance.

Notice, that the intra-state MTM tends to be close to diagonal (in average). This means, that when staying in a given state of an HMM, the same component of the density is likely to be dominating for the adjacent frames.

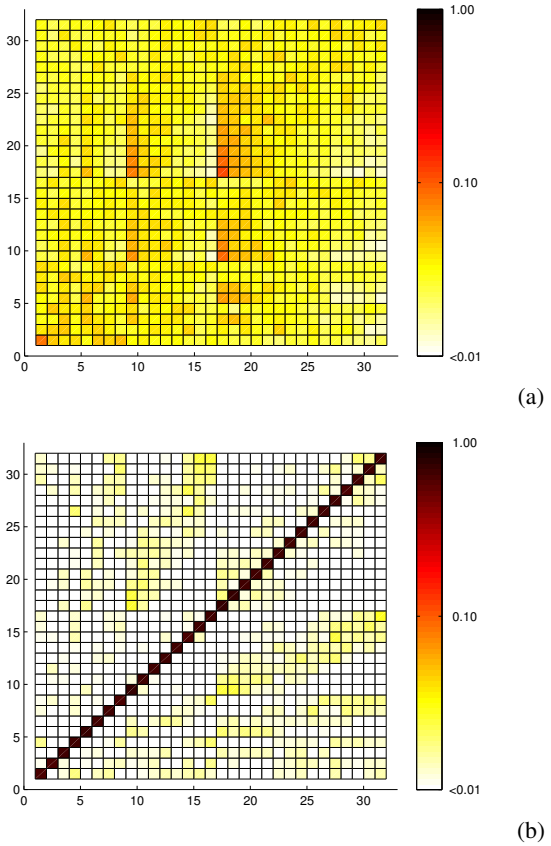

Figure 2: Inter-state(a) and intra-state(b) mixture transition matrices, averaged over states for CHiME data

The average value of the diagonal elements for intra-state MTM is 0.67 ± 0.02 for noisy CHiME and 0.58 ± 0.03 for TIdigits data. As it was noticed in [9], the distribution is also sparse.

Both sparsity and such sharp distribution for intra-sate transitions make training and decoding problems challenging. In fact, it is further shown, that sharp intra-state transitions constrain the trajectory and can lead to accuracy degradation, when the data contains non-stationary noise.

4.3. Experimental results on CHiME data

The results of the first set of experiments with MFCC features, derived from noised CHiME data, are summarized in table 3. The SGMM is initialized from the baseline SI system and re-estimated. For this set of experiments it was observed, that sharp diagonal distribution of intra-state MTMs (loops) significantly hurts the performance of the recognizer (this corresponds to "Loop MTM - trained" row in table 3).

Model		-6dB	-3dB	0dB	3dB	6dB	9dB	AVG
GMM baseline		**55.75**	**60.08**	69.58	77.67	80.08	84.25	**71.24**
SGMM training	Loop MTM							
MTM	trained	53.75	58.92	67.50	75.26	79.75	84.17	69.89
MTM	fixed	**57.08**	**61.08**	69.17	77.25	80.00	85.17	**71.63**
MTM+μ+σ	fixed	**57.83**	**62.08**	69.58	77.33	80.17	85.17	**72.03**

Table 3: Keyword recognition accuracy (%) for Dev set of CHiME 2013. Features from noisy data

A simple "work around" approach is to keep MTMs for state loops uniform (not re-estimate them). When intra-state MTM is forced to be uniform, SGMM outperforms GMM with larger gains in the noisy part (corresponds to the rows "Loop MTM - fixed"). Further improvement is achieved for this dataset when means and variances are jointly re-estimated with MTM.

Finally, the same experiment is reproduced with enhanced features, calculated with FASST [19]. The training is done in the same way, as for the SGMM trained from noisy data. The performances of the baseline GMM, of the SGMM with and without intra-state MTMs re-estimation and of the SGMM with means and variances re-estimated are summarized in Table 4.

Model		-6dB	-3dB	0dB	3dB	6dB	9dB	AVG
GMM baseline		73.00	78.00	82.92	85.83	89.17	90.33	**83.21**
SGMM training	Loop MTM							
MTM	trained	72.67	76.83	81.00	86.33	88.33	90.33	82.58
MTM	fixed	73.17	78.33	82.58	86.50	89.25	90.67	**83.42**
MTM+μ+σ	fixed	73.50	79.00	82.83	86.58	89.67	90.92	**83.75**

Table 4: Keyword recognition accuracy (%) for Dev set of CHiME 2013. Features from enhanced data

Overall, the relative improvements, achieved with SGMM on noisy and enhanced data, are similar. To verify the statistical significance of the results, McNemar test was done [22]. The test consists in analyzing the errors, produced by two systems and computing the probability P of how likely the improvement was done by chance. Comparing GMM and SGMM in the experiments with noisy features (Table 3) gives $P = 0.017$ and with enhanced features (Table 4) $P = 0.040$, which means that the results are statistically significant (with respect to 95% confidence interval).

5. Discussion

Tables 2, 3 and 4 summarize the improvements, which are achieved by SGMM compared to HMM-GMM.

Decoding clean read speech (as in TIdigits) imposes troubles for conventional HMM if the data comes from the speakers of different age and gender. SGMM shows significant improvements in such conditions. Re-estimating only MTM (row "SGMM: MTM" in Table 2) improves the WER by about 30% relative for both child and adult data. This leads to the conclusion, that temporal dependencies (explicit trajectories) introduced by SGMM are useful for modeling speaker variability.

Different observations are drawn from the experiments with noisy data. Training with noisy (Table 3), or enhanced (Table 4) data demonstrates similar behavior.

First, for CHiME data intra-state transitions hurt the model accuracy (rows "Loop MTM-trained" in tables 3 and 4). Interestingly, that at the same time the likelihood of the training data is much higher, when both inter- and intra-state MTMs are re-estimated. This leads to the idea, that SGMMs tend to over-fit, which leads to the degradation if the acoustic mismatch between training and testing data is significant.

Second, after fixing (not re-estimating) intra-state MTMs the improvement is not as large, as for TIdigits, but still statistically significant. Overall, about 2% relative improvement is achieved for the full CHiME data set. Up to 5% is also achieved for -6 and -3 dB subsets with noisy MFCC.

6. Conclusion

This work has described an experimental study of Stranded GMM, which is an extension of HMM-GMM model with additional temporal dependencies between components of the observation densities.

Three types of signal variability were studied. First, gender and age in clean speech were investigated. Then, the same experiments were carried on speech corrupted by non-stationary noise and on enhanced speech with residual noise.

Although SGMM was originally proposed for robust speech recognition of noisy data, in this paper it was demonstrated, that it provides the largest improvement for clean speech with adult and children data. When the signal is corrupted by non-stationary noise, SGMM improves the accuracy not as greatly and only if intra-state MTMs are not re-estimated.

Although they require increased computation power, such extended models, as SGMM, are certainly interesting for the future research. One research direction is the introduction of speaker, or noise information in the SGMM by defining some SGMM parameters from pre-clustered data, or by involving speaker adaptation. Finally, the issue with dominating diagonal probabilities of intra-state MTMs in CHiME data opens new questions, which should lead to improved re-estimation algorithm for MTM involving sparsity constraints, or smoothing.

7. References

[1] I. Illina and Y. Gong, "Elimination of trajectory folding phenomenon: HMM, trajectory mixture HMM and mixture stochastic trajectory model," in *Proc. ICASSP*. IEEE, 1997, vol. 2, pp. 1395–1398.

[2] L. Deng, "Dynamic Speech Models: Theory, Algorithms, and Applications," *Synthesis Lectures on Speech and Audio Processing*, vol. 2, no. 1, 2006.

[3] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 12, pp. 1857–1869, 1989.

[4] Y. Gong and J.-P. Haton, "Stochastic trajectory modeling for speech recognition," in *Proc. ICASSP*. IEEE, 1994, vol. 1, pp. I–57.

[5] W. Goldenthal and J. Glass, *Statistical trajectory models for phonetic recognition*, Ph.D. thesis, Massachusetts Institute of Technology, Dept. of Aeronautics and Astronautics, 1994.

[6] J. Bilmes, "Buried Markov models for speech recognition," in *Proc. ICASSP*. IEEE, 1999, vol. 2, pp. 713–716.

[7] F. Korkmazskiy, B.-H. Juang, and F. Soong, "Generalized mixture of HMMs for continuous speech recognition," in *Proc. ICASSP*. IEEE, 1997, vol. 2, pp. 1443–1446.

[8] G. Hinton, L. Deng, D. Yu, G. Dahl, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[9] Y. Zhao and B.-H. Juang, "Stranded Gaussian mixture hidden Markov models for robust speech recognition," in *Proc. ICASSP*. IEEE, 2012, pp. 4301–4304.

[10] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW*, 2000.

[11] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

[12] R. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 34–43, 2012.

[13] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[14] P. Zhan and A. Waibel, "Vocal tract length normalization for large vocabulary continuous speech recognition," in *Technical report*, DTIC Document, 1997.

[15] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second'CHiME'Speech Separation and Recognition Challenge: Datasets, tasks and baselines," in *Proc. ICASSP*, 2013, pp. 69–74.

[16] D. Tran, E. Vincent, and D. Jouvet, "Extension of uncertainty propagation to dynamic MFCCs for noise robust ASR," in *Proc. ICASSP (to appear)*, 2014.

[17] "http://cmusphinx.sourceforge.net," 2013.

[18] A. Ozerov, E. Vincent, and F. Bimbot, "A General Flexible Framework for the Handling of Prior Information in Audio Source Separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 4, pp. 1118 – 1133, 2012.

[19] A. Ozerov, E. Vincent, et al., "Using the FASST source separation toolbox for noise robust speech recognition," in *International Workshop on Machine Listening in Multisource Environments (CHiME)*, 2011.

[20] R. Leonard and G. Doddington, "Tidigits speech corpus," *Texas Instruments, Inc*, 1993.

[21] D. C. Burnett and M. Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in *Proc. ICSLP*. IEEE, 1996, vol. 2, pp. 1145–1148.

[22] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP*. IEEE, 1989, pp. 532–535.